محمد صادق عبدالهی[1]، دکتر عبدالرسول جوکار[2]

تاریخ دریافت: ۸۶/۱۲/۵
تاریخ پذیرش: ۸۸/۷/۲۱

چالش‌های شیوه نگارش زبان فارسی در بازیابی اطلاعات از موتورهای کاوش وب

چکیده

شیوه‌های گوناگون در نوشتن یک واژه، یکی از مشکلاتی است که در بازیابی مدارک مربوطه از طریق ابزارهای جستجو وجود دارد. در زبان فارسی نیز برخی از واژه‌ها به ریخت‌های متفاوتی نوشته می‌شوند. به همین دلیل این پژوهش به بررسی مسائلی پرداخته است که کاوشگران فارسی در کاوش ریخت‌های مختلف یک واژه با آن روبه‌رو هستند. برای پاسخگویی به سؤال‌های پژوهش از دو روش پیماش مقایسه‌ای و اسنادی استفاده شده است. جامعه پژوهش شامل سه موتور کاوش گوگل، یاهو و آلتاویستا است که امکان جستجو به زبان فارسی را فراهم کرده‌اند.

با مطالعه و دقت در متون فارسی، سیاهه‌ای شامل ۱۷ کلیدواژه انتخاب شد که هرکدام نمایان‌گر یک مورد از چالش‌های زبان فارسی در بازیابی اطلاعات هستند. پژوهشگران کلیدواژه‌ها را وارد فیلد جستجوی ابزارهای کاوش انتخابی کرده و تعداد بازیافت‌ها برای هرکدام از موتورهای کاوش را ثبت کردند.

تجزیه و تحلیل داده‌ها در دو سطح آمار توصیفی و استنباطی انجام شد. یافته‌ها نشان داد که موتورهای کاوش وب، شیوه‌های نگارش زبان فارسی را به منظور بهبود کاوش، مورد توجه قرار نداده‌اند. همچنین رابطه معناداری بین شکل واژه و نوع ابزار جستجو وجود دارد.

کلیدواژه‌ها: *بازیابی اطلاعات، موتورهای کاوش، شیوه نگارش، زبان فارسی*

۱ کارشناس ارشد کتابداری و اطلاع‌رسانی sadeqabdolahi@yahoo.com
۲ عضو هیأت علمی دانشگاه شیراز jowkar2003@yahooo.com

مقدمه

در زمانه‌ای به سر می‌بریم که در آن اطلاعات به شکل الکترونیکی تولید شده و انتقال می‌یابد. رشد علمی، فنی و فرهنگی در گرو برقراری ارتباط زبانی و کلامی با دنیای الکترونیکی دانش و فرهنگ است که وب نام دارد و این جز با تقویت کیفی زبان ملی میسر نمی‌شود. لیکن زبان فارسی، در تلاقی با جهان الکترونیکی، به‌ویژه از بعد شیوه نگارش[1]، دارای معضلاتی است که کاوش در محتویات آن را دچار مشکل می‌کند. این مقوله مستلزم تمهیداتی چند است تا زبان فارسی را از زبان شعر و ادب و عرفان، به زبانی مناسب با پهنه الکترونیکی دادوستد دانش، تبدیل کند (صدیقی، ۱۳۸۳).

پیرایش بر روی زبان‌های دنیا خیلی پیشتر از این آغاز شده است. گسترش استانداردهای آماده‌شده برای زبان، ساده‌کردن و کم کردن قاعده‌های پیچیده و استثناها در زبان روزمره و یکسان‌کردن گفتار و نوشتار از کارهایی است که بر روی بسیاری از زبان‌ها انجام شده‌است. برای نمونه، در زبان انگلیسی، زبان‌شناسان بسیاری از قاعده‌های این زبان را پیراسته‌اند و یادگیری این زبان را ساده کرده‌اند (یوسفان نجف آبادی، ۱۳۸۲).

در حال حاضر وبلاگ‌[2]های ایرانی از نظر فراوانی، جزو پنج کشور برتر جهان هستند. اما ویژگی-های منفی آنها به این شرح است:

۱- شیوه نگارش وبلاگ‌ها بیشتر غیراستاندارد و متغیر است.

۲- نوشته‌های وبلاگ‌ها به نسبت حاوی غلط‌های املایی و نگارشی زیادی است، هرچند که بیشتر وبلاگ‌های مهم و پرخواننده، نگارش قابل قبولی دارند.

۳- شیوه نگارش وبلاگ‌ها، تابع محدودیت‌های محیط الکترونیکی و عدم تطبیق آن با الزام‌های خط فارسی است.

از طرف دیگر، همه نویسندگان وبلاگ‌ها، بنا به اهمیتی که زبان فارسی به عنوان زبان رسمی ما دارد و به حکم مسئولیتی که به عنوان صاحب رسانه دارند، باید خود را موظف بدانند که برای حفظ سلامت زبان فارسی در رسانه خود تلاش کنند. برای این منظور لازم است نویسندگان وبلاگ‌ها، بر کاربرد

---

[1] رسم‌الخط

[2] Weblog

زبان در وبلاگ خود نظارت کنند و استانداردهای نگارشی زبان و خط فارسی را رعایت کنند.

مسئله و اهمیت پژوهش

در قرن بیست‌ویکم اطلاعات به سرعت به سمت رقومی شدن پیش می‌رود. وب، بزرگترین مرجع اطلاعات در عصر ما محسوب می‌شود. اما شیوه نگارش فارسی باعث بروز چالش‌های جدی در امر نمایه‌سازی این زبان شده است. مسائل مربوط به خط، یکی از جنبه‌های مهم برنامه‌ریزی زبان[1] است. ایجاد خط، انتخاب خط مشترک، تغییر خط و اصلاح آن از رایج‌ترین شکل‌های برخورد با خط به حساب می‌آید. گفتار و نوشتار دو بستر برای تحقق زبان هستند که گفتار، بازتاب طبیعی‌تر آن است، اما نوشتار به دلیل آنکه صورت ثابتی دارد و تحولات زبان را منعکس نمی‌کند، نیاز به برنامه‌ریزی زبانی و اصلاح دارد (اسلامی، ۱۳۸۱). چنانچه الگویی مناسب برای رفع این چالش‌ها ارائه شود، بخشی از مسائلی که امروز گریبان‌گیر زبان فارسی است، رفع خواهد شد. از جمله این مسائل می‌توان به موارد زیر اشاره کرد: جستجوی بهینه در وب، ایجاد پایگاه‌های اطلاعاتی به زبان فارسی، ایجاد نظام هم‌آهنگ اطلاع‌رسانی در کشور و مسائل دیگر (حری، ۱۳۷۲).

امروزه روش غالب در جستجوی اطلاعات از موتورهای کاوش وب، روش کلیدواژه‌ای است. اما جستجو به این روش، دشواری‌های خاص خود را دارد. چنانچه فردی به دنبال اطلاعاتی در مورد "آب گرمکن" باشد، این کلیدواژه را می‌تواند به چهار شکل بنویسد: "آب گرم کن، آبگرم کن، آب گرمکن و آبگرمکن". بنابراین موتورهای کاوش وب، برای هرکدام از این شکل‌ها، تعداد بازیافت‌های متفاوتی بازیابی خواهند کرد. چنانچه کاربری تنها یک شکل از این چهار مورد را به‌کار ببرد، اطلاعاتی که به اشکال دیگر نوشته شده است را از دست خواهد داد. سؤال مهمی که در اینجا مطرح می‌شود این است که: چگونه می‌توان بر این مسئله فایق آمد؟ آیا باید دست به اصلاح شیوه نگارش فارسی زد، یا اینکه نظام‌هایی پیشرفته و سازگار با این شیوه نگارش طراحی کرد؟

پیشینه پژوهش

ابزارهای جستجو اساسا بر مبنای زبان انگلیسی طراحی شده‌اند و کشورهای غیر انگلیسی‌زبان چالش‌های مشابهی با آن‌ها دارند. در پژوهش‌های انجام شده در خارج از کشور، پژوهشگران به مقایسه و

[1] Language Planning

ارزیابی موتورهای کاوش بین‌المللی و محلی پرداخته‌اند. ریشه‌سازی[1]، کوتاه‌سازی[2] و جستجوی مترادف‌ها از جمله معیارهایی است که این پژوهشگران از آنها در ارزیابی‌های خود به کار برده‌اند. اما از آن جایی که بسیاری از این امکانات در ابزارهای کاوش فارسی وجود ندارد، پژوهشگران ایرانی را به سمت ارزیابی‌های متفاوت‌تری سوق داده است. پژوهشگران ایرانی معیارهایی مانند عملگرهای بولی، میزان پیوند به یک موتور کاوش، حجم پایگاه اطلاعاتی، رتبه بندی بازیافت‌ها، نمایش اطلاعات، واسط کاربر، روز آمد بودن اطلاعات، سرعت بازیابی اطلاعات، نمایه سازی اطلاعات را در پژوهش های خود به کار برده‌اند.

الف) پیشینه پژوهش در خارج از ایران

هدلاند و دیگران (Hedlund et al., 2000) به بررسی ویژگی های زبان سوئدی از دیدگاه بازیابی اطلاعات پرداختند. مشکلی که این پژوهشگران با آن مواجه بودند بازیابی ضعیف اطلاعات به زبان سوئدی بود. این زبان، ویژگی‌های منحصربه‌فردی دارد. از آن‌جمله می‌توان به مذکر و مؤنث بودن نام‌ها و نیز فراوانی استفاده از واژه‌های هم‌نگاشت اشاره کرد. آنها مطالعه‌ای مقایسه‌ای بر روی زبان‌های سوئدی، فنلاندی و انگلیسی انجام دادند تا میزان ابهام های واژگانی در این زبان‌ها را شناسایی کنند. پژوهشگران پیشنهاد می‌کنند که برچسب‌گذاری ادات سخن[3] برای بازیابی واژه‌های هم‌نگاشت، می‌تواند مفید باشد.

سروکا (Sroka, 2000) نسخه‌های لهستانی چند ابزار جستجوی بین‌المللی را به همراه چند موتور کاوش محلی مورد سنجش قرار داد. مهم‌ترین معیار این سنجش، دقت[4] ابزار کاوش بود که بر اساس ربط قضاوتی ۱۰ نتیجه نخست هر کاوش محاسبه شد. پژوهشگر تعداد بازیافت‌ها و زمان صرف شده برای یک کاوش را در مورد هر کدام از موتورهای کاوش ثبت کرد. در نتیجه این پژوهش "پالسکی اینفوسیک[5]" به عنوان بهترین ابزار کاوش انتخاب شد.

مونز و دوریجکه (Monz & De Rijke, 2002) با تمرکز روی اثرات تحلیل‌های ریخت‌شناسی همچون ریشه‌سازی و جداسازی واژه‌های مرکب، به بررسی کارآیی بازیابی اطلاعات پرداختند. این مطالعه برروی زبان‌های هلندی، آلمانی و ایتالیایی انجام شد. یافته‌ها نشان داد که بازیابی اطلاعات در حدود ۲۵٪ برای زبان آلمانی، ۶۹٪ برای زبان هلندی و ۲۵٪ برای زبان ایتالیایی بهبود داشته است.

[1] Stemming
[2] Truncation
[3] Part of speech
[4] Precision
[5] Polski Infoseek

بارایلان و گتمان (Bar-Ilan & Gutman, 2002) توانایی ابزارهای جستجو را در مورد زبان‌های غیرانگلیسی مورد بررسی قرار دادند. ۴ زبان روسی، فرانسوی، مجاری و عبری جامعه این پژوهش به شمار می‌آیند. برای هر کدام از این زبان‌ها، ۳ موتور کاوش عمومی یعنی "آلتاویستا[1]"، "فست[2]" و "گوگل[3]" به همراه چند موتور کاوش محلی (مخصوص هر کدام ازاین زبان‌ها) آزمایش شد. این بررسی نشان داد که موتورهای کاوش عمومی وب، ویژگی‌های زبانی زبان‌های غیر انگلیسی را در جستجوی اطلاعات نادیده می‌گیرند.

مقداد (Moukdad, 2005) در پژوهشی، عملکرد ۳ ابزار جستجوی عمومی را با ۳ موتور کاوش عربی (که به طور خاص مسائل زبان‌شناختی عربی را لحاظ می‌کنند) مورد مقایسه قرار داد. یافته‌ها نشان داد که موتور های کاوش عمومی، نظیر "آلدوب[4]"، "آلتاویستا" و "گوگل" در بازیابی مدارك عربی، ناقص عمل می‌کنند. همچنین یافته‌های این پژوهش، نیاز به پژوهش‌های بیشتر در زمینه عملی بودن ابزارهای جدید بازیابی اطلاعات در موتور های کاوش را نشان داد.

تاث (Toth, 2006) به بررسی قابلیت‌های زبان‌شناختی موتور های کاوش انگلیسی و مجاری پرداخت. پژوهشگر ۳ ابزار جستجوی انگلیسی "گوگل"، "آلتاویستا" و "آلدوب" را با ۵ موتور کاوش محلی مورد مقایسه قرار داد. تحلیل داده‌ها بر پایه چند شاخص انجام شد که عبارت بودند از: ریشه‌سازی، بازیابی لهجه‌های مختلف، کوتاه‌سازی و جستجوی مترادف‌ها. یافته‌ها حاکی از آن بود که موتور های کاوش محلی، مسائل زبان مجاری را بهتر از موتور های کاوش انگیسی مورد توجه قرار داده بودند. ابزارهای انگلیسی‌زبان، لهجه‌های مختلف زبان مجاری را به خوبی پشتیبانی نمی‌کردند، که این امر منجر به بازیابی ضعیف اطلاعات می‌شد.

ب) پیشینه در ایران

کوشا (۱۳۸۱) با استفاده از معیارهای مستند بـه ارزیـابی‌جداگانـه و نیـز تجزیـه و تحلیـل مقایـسه‌ای ابزارهای کاوش دارای واسط جستجوی فارسی‌پرداخـت. شـش ابـزار کـاوش برگزیـده از طریـق ۲۷ معیـار مرتبط با قابلیت‌های جستجو و بازیابی اطلاعات با یکدیگر مورد مقایسه قـرار گرفتنـد. موتورهـای کـاوش

[1] Altavista
[2] Fast
[3] Google
[4] All the web

انتخابی عبارت بودند از: "گوگل"، "ایران‌کلیک[1]"، "ایران‌هو[2]"، "ایران‌مهر[3]"، "پارسیک[4]"، و "اپن‌دایرکتوری[5]". نتیجه پژوهش نشان داد که از نظر امکانات جستجو و بازیابی اطلاعات، ابزار کاوش "گوگل" در رتبه نخست و راهنمای موضوعی "ایران‌هو" در رتبه دوم قرار دارند. به منظور بررسی عامه‌پسند و رایج بودن ابزارهای کاوش، پژوهشگر تعداد صفحه‌ها یا وب سایت پیوند داده شده به آنها را مورد مقایسه قرار داد. نتیجه نشان داد که رابطه مستقیمی میان رایج و عامه‌پسندبودن ابزارهای کاوش مورد مطالعه با توانایی‌های جستجوی اطلاعات آنها وجود ندارد.

یوسفان نجف آبادی (۱۳۸۲) در پژوهشی با عنوان "یک نظام بازیابی متنی برای زبان فارسی برپایه معانی پنهان"، نظامی را طراحی کرده است که با استفاده از نمایه‌گذاری معانی به بازیابی اطلاعات متنی زبان فارسی می‌پردازد. کارآیی این نظام با ریشه‌یابی و بدون ریشه‌یابی با استفاده از یک مجموعه اسناد گردآوری‌شده به این منظور و به کمک معیارهای دقت و بازیافت مورد ارزیابی قرار گرفته‌است. برای کمک به یافتن فهرست واژه‌های سراسری و ریشه‌یابی، یک زبان برنامه‌نویسی ساده، طراحی و بر پایه قاعده‌های زبان فارسی، روشی نوین برای شناسایی خودکار فعل‌های فارسی پیشنهاد شد.

راثی ساربانقلی (۱۳۸۴) در پژوهش خود به بررسی مشکلات جستجو و بازیابی اطلاعات به زبان فارسی در اینترنت به کمک کاربران مرکز اینترنت دانشگاه آزاد اسلامی شبستر پرداخت. یافته های پژوهش نشان داد که ۷۷٪ کاربران از جستجوی پیشرفته گوگل استفاده می‌کنند. بیشتر مشکل کاربران در جستجو، عدم توجه ایشان به شکل‌های مختلف نوشتاری واژه و عدم استفاده از عملگر "OR" بود. این پژوهش، رابطه‌ی معناداری بین گذراندن دوره‌های آموزشی و نیز مدت استفاده کاربران از اینترنت با مهارت آنها را نشان داد.

مطالعه پژوهش‌های انجام شده نشان داد که موتورهای کاوش در بازیابی منابع بر اساس شکل‌های مختلف یک واژه توانمند نیستند. با توجه به آنچه در مقدمه و پیشینه پژوهش بحث شد، از آنجایی که بسیاری از مدارک که با املاءهای مختلف یک کلمه در محیط اینترنت وجود دارند، نمی‌توانند توسط موتورهای کاوش بازیابی شوند، ضروری است که از طریق پژوهش بتوان راهکارهای لازم را شناسایی

[1] Iran click
[2] Iranhoo
[3] Iran mehr
[4] Parseek
[5] Open directory

کرد. بنابراین هدف این پژوهش شناسایی راه‌ها و روش‌هایی است که از مشکلات بازیابی مدارک که بر اساس شیوه نگارش واژه‌ها به وجود می‌آید جلوگیری کرد. در این راستا سؤال‌های پژوهشی زیر طراحی شده است:

۱. کدام‌یک از ویژگی‌های شیوه نگارش زبان فارسی در بازیابی اطلاعات از وب مشکل ایجاد می‌کند؟

۲. آیا ابزارهای کاوش بین‌المللی (با در نظر گرفتن شیوه نگارش فارسی) نتایج جستجوی یکسانی برای شکل‌های مختلف یک کلمه به دست‌می‌دهند؟

۳. آیا رابطه معناداری بین شکل واژه‌ها و نوع ابزار جستجو وجود دارد؟

طرح پژوهش

برای پاسخگویی به سؤال نخست پژوهش، روش اسنادی انتخاب گردید. برای پاسخگویی به سؤال های دوم و سوم پژوهش از روش پیماش مقایسه‌ای استفاده شد. با کمک روش پیش گفته مشکلاتی که شکل‌های مختلف واژه در ابزارهای کاوش انتخابی ایجاد می‌کنند، مورد بررسی و تجزیه و تحلیل قرار گرفت.

بررسی ادبیات پژوهش نشان داد که هفت موتور کاوش بین‌المللی یعنی آلتاویستا، اکسایت[۱]، گوگل، هات بات[۲]، اینفوسیک[۳]، لایکاس[۴] و یاهو[۵] به عنوان پر استفاده‌ترین ابزارهای کاوش در دنیا شناخته شده‌اند. در این میان تنها "گوگل"، "یاهو" و "آلتاویستا" امکان جستجو به زبان فارسی را فراهم کرده‌اند. بنابراین، این سه موتور کاوش به عنوان بستر برای پیشبرد این پژوهش در نظر گرفته شدند. ابزارهای کاوش بین‌المللی، از عنکبوت[۶] یا خزنده به منظور شناسایی و نمایه‌سازی صفحه‌ها یا سایت‌های وب در زبان‌های مختلف از جمله زبان فارسی استفاده می‌کنند. این روش نوعی نمایه‌سازی خودکار می‌باشد و می تواند صفحه‌های فارسی را در قالب یونی‌کد[۷] شناسایی و در پایگاه خود ذخیره کنند.

[1] Excite
[2] Hotbot
[3] Infoseek
[4] Lycos
[5] Yahoo
[6] Spider
[7] Unicode

با بررسی پژوهش‌های فارسی، سیاهه‌ای شامل ۱۷ کلیدواژه به صورت تعمدی به عنوان نمونه انتخاب شد. این کلیدواژه‌ها هر کدام نمایانگر یک مورد از چالش‌های زبان فارسی در بازیابی اطلاعات هستند. این کلیدواژه‌ها که به عنوان وسیله گردآوری داده‌ها شناخته می‌شوند، عبارتند از:

موسی یا موسا	اتاق یا اطاق
املاء یا املا	توراه یا تورات
باغها یا باغ ها	عطایی یا عطائی
موّحدی یا موحدی	مورچه گان یا مورچگان
پتاسیم یا پتاسیوم	خانه من یا خانه‌ی من
زبان شناس یا زبانشناس	مسؤول یا مسئول
مسئله یا مسأله	دقیقاً یا دقیقن
شمشیرباز یا شمشیرِباز	پرتو آفتاب یا پرتوی آفتاب
	فرایند یا فرآیند

به منظور انجام کاوش‌ها، نخست به بخش جستجوی پیشرفته ابزار کاوش وارد شده، سپس در قسمت زبان‌ها، زبان فارسی به عنوان پیش‌فرض جستجو انتخاب شد. هر یک از شکل‌های مختلف واژه را وارد فیلد جستجوی ابزارهای کاوش انتخابی کرده، سپس تعداد یافته های حاصل از جستجو توسط هر یک از موتور های کاوش ثبت شد. اگرچه موتور کاوش "آلتاویستا"، امکان جستجو به زبان فارسی را فراهم کرده‌است، اما در قسمت جستجوی پیشرفته، امکان انتخاب زبان به کاربران داده نمی شود. این مشکل باعث شد که به هنگام بازیابی برخی از کلیدواژه‌ها، تعدادی از بازیافت‌ها به زبان عربی ارائه شود. مشخص است که صافی زبان[1] در اینجا درست عمل نکرده‌است.

تجزیه وتحلیل داده ها

در این پژوهش، تجزیه و تحلیل داده‌ها در دو سطح توصیفی و استنباطی انجام شد. توصیف داده‌ها با آماره‌های توصیفی (جدول فراوانی و نمودار درصد) صورت گرفت. برای بررسی معناداری

[1] Language filter

رابطه بین نوع موتور کاوش و شکل واژه از آماره‌های استنباطی (آزمون خی دو[1] و ضریب فی[2]) استفاده شده است.

سؤال ۱. کدامیک از ویژگی‌های شیوه نگارش زبان فارسی در بازیابی اطلاعات از وب مشکل ایجاد می‌کند؟

در پاسخ به این سؤال باید به طور کلی شرایطی که موجب نگارش یک واژه به ریخت‌های گوناگون می‌شود را مورد مطالعه قرار داد. برخی از مطالعه‌های صورت گرفته (مانند مرتضایی، ۱۳۷۶) نیز به این امر توجه کرده‌اند. نتیجه بررسی نشان داد که عامل‌هایی مانند مهارت استفاده از حالت‌های نوشتاری یا دستوری سایر زبان‌ها مانند عربی، سلیق‌های مختلف در نگارش واژه‌ها و یا برگردان آنها به فارسی صورت‌های مختلف یک واژه را در نگارش تشکیل می‌دهد. نمونه‌هایی در این رابطه به شرح زیر می‌باشد:

۱. برگردان واژه‌های خارجی مانند پستالوزی/پستالزی، پتاسیم یا پتاسیوم
۲. نشانه‌های جمع مانند مدارس/ مدرسه‌ها، استادان/ استادها
۳. پیوسته‌نویسی یا جدانویسی مانند مردم شناسی/ مردمشناسی، روان‌شناسی/ روانشناسی
۴. تنوین مانند اصلاً/ اصلن، دقیقاً/ دقیقن
۵. کسره اضافه مانند اسبِ سواری/ اسبْ سواری
۶. صامت میانجی "ی" مانند پرتوی آفتاب/ پرتو آتاب
۷. همزه برای واژه‌های مختوم به "ها"ی بیان حرکت مانند جامه من/ جامهٔ من/ جامه‌ی من
۸. همزه پایانی مانند انشاء/ انشا، املاء/ املا
۹. همزه پایانی متصل به "یا"ی وحدت یا نکره مانند کرسی "یـ"/ کرسی "ئـ" (عطایی/ عطائی)
۱۰. به کار بردن همزه به صورت‌های مختلف مانند مسئله/ مسأله، مسئول/ مسؤول

[1] χ^2
[2] Phi coefficient

۱۱. الف مقصوره که گاه به صورت الف مانند اسماعیل، هارون و جز آن نوشته می شود و گاه به همان شکل عربی مانند عیسی

۱۲. تنوع املایی مانند پتر کبیر/ پطر کبیر، اختاپوس/ اختاپوث

۱۳. استفاده از "ا" و "آ" به جای یکدیگر مانند فرآهم/ فراهم، برآیند/ برایند

۱۴. کلمات خاص در پیوسته‌نویسی و جدانویسی مانند علاقمند/ علاقه‌بندی، اندیشمند/ اندیشه‌مند

۱۵. "تای" منقوط مانند صلاۀ/ صلات، مشکوۀ/ یا مشکات

۱۶. نشانه تشدید مانند معین/ معیّن، علی/علّی

سؤال ۲. آیا ابزارهای کاوش بین‌المللی (با در نظر گرفتن شیوه نگارش فارسی) نتایج جستجوی یکسانی برای شکل‌های مختلف یک کلمه به دست می‌دهند؟

۱. شیوه برگردان واژه‌های خارجی

جدول ۱. آمار بازیافت‌ها برای مقوله شیوه برگردان واژه‌ها خارجی (n=2)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	یاهو	آلتاویستا	
شکل واژه	پتاسیم	فراوانی	۱۰۴۰۰۰	۲۳۹۰۰	۲۵۴۰۰	۱۵۳۳۰۰
		درصد	٪۶۷/۸	٪۱۵/۶	٪۱۶/۶	٪۱۰۰
	پتاسیوم	فراوانی	۹۳	۳۴۰	۳۴۲	۷۷۵
		درصد	٪۱۲	٪۴۳/۹	٪۴۴/۱	٪۱۰۰
جمع کل		فراوانی	۱۰۴۰۹۳	۲۴۲۴۰	۲۵۷۴۲	۱۵۴۰۷۵
		درصد	٪۶۷/۶	٪۱۵/۷	٪۱۶/۷	٪۱۰۰

همان‌طورکه در جدول نشان داده شد، به طور کامل دو نتیجه متفاوت برای این دو کلیدواژه بازیابی شده است. اطلاعات بیشتری با کلید واژه "پتاسیم" ذخیره شده است و انتخاب کلیدواژه "پتاسیوم" باعث از دست رفتن این اطلاعات می‌شود. همچنین در هیچ‌یک از موتور های کاوش، تمهیدی برای

بازیابی صفحه های دارای کلیدواژه "پتاسیم" در هنگام جستجوی کلیدواژه "پتاسیوم" اندیشیده نشده است.

۲. شیوه نگارش نشانه‌های جمع

جدول ۲: تعداد بازیافت‌ها برای مقوله شیوه نگارش نشانه‌های جمع (n=2)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	یاهو	آلتاویستا	
شکل واژه	باغ ها	فراوانی	۴۴۰۰	۲۷۰۰	۲۸۵۰۰	۳۵۶۰۰
		درصد	٪۱۲/۴	٪۷/۶	٪۸۰/۱	٪۱۰۰
	باغها	فراوانی	۶۴۴۰۰	۱۶۶۰۰	۱۹۰۰	۸۲۹۰۰
		درصد	٪۷۷/۷	٪۲۰	٪۲/۳	٪۱۰۰
جمع کل		فراوانی	۶۸۸۰۰	۱۹۳۰۰	۳۰۴۰۰	۱۱۸۵۰۰
		درصد	٪۵۸/۱	٪۱۶/۳	٪۲۵/۷	٪۱۰۰

چناچه کاربری کلیدواژه "باغها" را انتخاب کند، بیشتر اطلاعات موجود که با کلیدواژه "باغ ها" ذخیره شده است را از دست می‌دهد. ازطرف دیگر انتخاب "باغ ها" نیز، ریزش کاذب به بار می‌آورد، چرا که موتورهای کاوش، هر فاصله خالی بین واژه‌ها را همچون عمل‌گر[۱] And در نظر می‌گیرد. بنابراین، صفحه‌هایی بازیابی می‌شوند که در آن کلمه "ها" به‌تنهایی آمده است. پس هیچکدام از موتورهای کاوش تمهیدی برای این مسئله نیندیشیده‌اند.

۳. پیوسته‌نویسی و یا جدانویسی ترکیب‌ها

جدول ۳. تعداد بازیافت‌ها برای مقوله پیوسته‌نویسی و یا جدانویسی ترکیب‌ها (n=3)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	یاهو	آلتاویستا	
شکل واژه	زبان شناس	فراوانی	۱۷۳۰۰۰	۲۰۲۰۰۰	۲۰۸۰۰۰	۵۸۳۰۰۰
		درصد	٪۲۹/۷	٪۳۴/۶	٪۳۵/۷	٪۱۰۰
	زبانشناس	فراوانی	۸۵۹	۹۴۲	۱۰۹۰	۲۸۹۱

[1] Operator

متغیر		آماره	گوگل	یاهو	آلتاویستا	جمع کل
		درصد	٪۲۹/۷	٪۳۲/۶	٪۳۷/۷	٪۱۰۰
	زبان‌شناس	فراوانی	۸۵۹	۲۰۲۰۰۰	۲۰۸۰۰	۲۲۳۶۹۵
		درصد	٪۰/۴	٪۹۰/۳	٪۹/۳	٪۱۰۰
جمع کل		فراوانی	۱۷۴۷۱۸	۴۰۴۹۴۲	۲۲۹۸۹۰	۸۰۹۵۵۰
		درصد	٪۲۱/۶	٪۵۰	٪۲۸/۴	٪۱۰۰

همان‌طورکه از جدول برمی‌آید، این ترکیب را به سه صورت می‌توان نوشت: جدانویس (زبان شناس)، پیوسته‌نویس (زبانشناس) و بی‌فاصله‌نویس (زبان‌شناس) (کابلی، **۱۳۷۳**). موتورهای کاوش "آلتاویستا" و "یاهو" برای شکل‌های جدانویس و بی‌فاصله‌نویس دو نتیجه یکسان به بار آورده‌اند، در حالی‌که موتور کاوش "گوگل" شکل پیوسته‌نویس و بی‌فاصله‌نویس را یکی محسوب کرده‌است. چنین عملکردی به روبات یا عنکبوت موتورهای کاوش برمی‌گردد که هرکدام طبق پیش‌فرض‌هایی که برای آنها تعریف شده است، واژه‌ها را شناسایی کرده و در پایگاه خود، نمایه می‌کنند. عملکرد مطلوب آن است که موتورهای کاوش برای هر سه شکل، نتیجه یکسانی به بار آورند.

۴. مشکل تنوین

جدول ۴: تعداد بازیافت‌ها برای مقوله تنوین (n=2)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	یاهو	آلتاویستا	
شکل واژه	دقیقاً	فراوانی	۴۶۵۰۰۰	۶۸۸۰۰۰	۸۷۵۰۰۰	۲۰۲۸۰۰۰
		درصد	٪۲۲/۹	٪۳۳/۹	٪۴۳/۱	٪۱۰۰
	دقیقن	فراوانی	۱۲۳۰۰	۲۱۸۰	۲۲۳۰	۱۶۷۱۰
		درصد	٪۷۳/۶	٪۱۳	٪۱۳/۳	٪۱۰۰
جمع کل		فراوانی	۴۷۷۳۰۰	۶۹۰۱۸۰	۸۷۷۲۳۰	۲۰۴۴۷۱۰
		درصد	٪۲۳/۳	٪۳۳/۷	٪۴۳	٪۱۰۰

در بعضی صفحه‌کلیدها، تنوین در جای همیشگی خود قرار نمی‌گیرد، این امر ماشین‌نویس‌ها را سر در گم می‌کند. به‌ناچار، تنوین نصب را که بیشتر در قیدها (مانند واقعاً، فوراً، جداً) کاربرد دارد به همان صورتی‌که خوانده می‌شوند (واقعن، فورن، جدن) به کار می‌برند. بر اساس نتیجه‌ایی که از جدول ۴ به

دست آمده است، می‌توان این‌طور استنباط کرد که این پراکندگی در کلیدواژه‌هایی که دارای تنوین هستند موجب از دست رفتن بخشی از اطلاعات موجود می‌گردد.

۵. کسره اضافه

جدول ۵: تعداد بازیافت‌ها برای مقوله کسره اضافه (n=2)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	**یاهو**	**آلتاویستا**	
شکل واژه	شمشیرباز	فراوانی	۱۶۳۰۰۰	۱۵۷۰۰۰	۱۵۰۰۰۰	۴۷۰۰۰۰
		درصد	٪۳۴/۷	٪۳۳/۴	٪۳۱/۹	٪۱۰۰
	شمشیرِباز	فراوانی	۷۱۷۰۰	۱۵۷۰۰۰	۱۵۰۰۰۰	۳۷۸۷۰۰
		درصد	٪۱۸/۹	٪۴۱/۵	٪۳۹/۶	٪۱۰۰
جمع کل		فراوانی	۲۳۴۷۰۰	۳۱۴۰۰۰	۳۰۰۰۰۰	۸۴۸۷۰۰
		درصد	٪۲۷/۷	٪۳۷	٪۳۵/۳	٪۱۰۰

در دو صورت، با وارد کردن یا نکردن نشانه کسره اضافه در نوشتار، موتور کاوش گوگل دو نتیجه متفاوت برای ترکیب **شمشیرباز** و **شمشیرِباز**، به دست می‌دهد، اما آلتاویستا و یاهو با آن یکسان عمل کرده‌اند.

۶. صامت میانجی "ی"

جدول ۶: تعداد بازیافت‌ها برای مقوله صامت میانجی "ی" (n=2)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	**یاهو**	**آلتاویستا**	
شکل واژه	پرتو آفتاب	فراوانی	۸۱۰۰۰	۷۸۸۰۰	۶۹۶۰۰	۲۲۹۴۰۰
		درصد	٪۳۵/۳	٪۳۴/۴	٪۳۰/۳	٪۱۰۰
	پرتوی آفتاب	فراوانی	۱۶۵۰	۱۵۷۰	۱۰۳۰۰	۱۳۵۲۰
		درصد	٪۱۲/۲	٪۱۱/۶	٪۷۶/۲	٪۱۰۰
جمع کل		فراوانی	۸۲۶۵۰	۸۰۳۷۰	۷۹۹۰۰	۲۴۲۹۲۰
		درصد	٪۳۴	٪۳۳/۱	٪۳۲/۹	٪۱۰۰

جوینده اطلاعات در حین کاوش، شاید هیچ‌گاه به این نکته نیندیشد که گذاشتن یا نگذاشتن میانجی "ی" چه تغییر زیادی در تعداد بازیافت‌ها خواهد گذاشت.

۷. استفاده یا عدم استفاده از "همزه"، برای واژه‌های مختوم به "های" بیان حرکت، در حالت مضاف

جدول ۷: تعداد بازیافت‌ها برای مقوله استفاده یا عدم استفاده از "همزه" (n=3)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	یاهو	آلتاویستا	
شکل واژه	خانه من	فراوانی	۲۸۰۰	۲۲۶۰	۹۲۱۰	۱۴۲۷۰
		درصد	٪۱۹/۶	٪۱۵/۸	٪۶۴/۵	٪۱۰۰
	خانهٔ من	فراوانی	۲۸۰۰	۲۲۶۰	۲۵۶۰	۷۶۲۰
		درصد	٪۳۶/۷	٪۲۹/۷	٪۳۳/۶	٪۱۰۰
	خانه‌ی من	فراوانی	۱۲۹۰۰۰۰	۱۲۳۰۰۰۰	۱۱۰۰۰۰۰	۳۶۲۰۰۰۰
		درصد	٪۳۵/۶	٪۳۴	٪۳۰/۴	٪۱۰۰
جمع کل		فراوانی	۱۲۹۵۶۰۰	۱۲۳۴۵۲۰	۱۱۱۱۷۷۰	۳۶۴۱۸۹۰
		درصد	٪۳۵/۶	٪۳۳/۹	٪۳۰/۴۵	٪۱۰۰

نتیجه غیر منتظره این است که ترکیب "خانه‌ی من"، بازیافت‌های بسیار بیشتری از دو ترکیب دیگر به دست داده است، در حالیکه نوشتن دو ترکیب دیگر، برای ماشین‌نویس آسان‌تر است.

۸. "همزه" پایانی

جدول ۸: تعداد بازیافت‌ها برای مقوله "همزه" پایانی (n=2)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	یاهو	آلتاویستا	
شکل واژه	املا	فراوانی	۳۴۳۰۰	۱۰۹۰۰	۴۹۱۰۰۰	۵۳۶۲۰۰
		درصد	٪۶/۴	٪۲	٪۹۱/۶	٪۱۰۰
	املاء	فراوانی	۹۷۷	۱۶۱۰	۲۱۴۰۰۰	۲۱۶۵۸۷
		درصد	٪۰/۵	٪۰/۷	٪۹۸/۸	٪۱۰۰
جمع کل		فراوانی	۳۵۲۷۷	۱۲۵۱۰	۷۰۵۰۰۰	۷۵۲۷۸۷
		درصد	٪۴/۷	٪۱/۷	٪۹۳/۷	٪۱۰۰

این ناهماهنگی در شیوه نگارش "همزه" پایانی، به راحتی موجب از دست دادن اطلاعات با ارزش می‌شود. به طور نمونه در موتور کاوش آلتاویستا، فردی که شکل دوم را به کار می‌برد، حدود ۲۸۰ هزار نتیجه جستجو را از دست می‌دهد.

۹. "همزه" پایانی متصل به "یای" وحدت یا نکره

جدول ۹: تعداد بازیافت‌ها برای مقوله "همزه" پایانی متصل به "یای" وحدت یا نکره (n=2)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	یاهو	آلتاویستا	
شکل واژه	عطائی	فراوانی	۱۸۴۰۰	۹۶۶۰۰	۱۰۹۰۰۰	۲۲۴۰۰۰
		درصد	٪۸/۲	٪۴۳/۱	٪۴۸/۷	٪۱۰۰
	عطایی	فراوانی	۸۲۳۰۰	۹۶۶۰۰	۱۰۹۰۰۰	۲۸۷۹۰۰
		درصد	٪۲۸/۶	٪۳۳/۶	٪۳۷/۹	٪۱۰۰
جمع کل		فراوانی	۱۰۰۷۰۰	۱۹۳۲۰۰	۲۱۸۰۰۰	۵۱۱۹۰۰
		درصد	٪۱۹/۷	٪۳۷/۷	٪۴۲/۶	٪۱۰۰

نکته جالب در این کاوش، این است که موتورهای جستجوی آلتاویستا و یاهو برای هردو شکل، نتیجه یکسانی داشته‌اند. اما مشخص نیست که چرا در واژه‌های دیگری که با همزه به کار می‌روند (مانند مسئله/ مسأله و مسئول/ مسؤول)، این اتفاق نیفتاده است.

۱۰. به کاربردن "همزه" به‌صورت های مختلف

جدول ۱–۱۰: تعداد بازیافت‌ها برای مقوله به کاربردن "همزه" به‌صورت های مختلف (n=2)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	یاهو	آلتاویستا	
شکل واژه	مسئول	فراوانی	۳۵۶۰۰۰	۱۵۳۰۰۰۰	۳۸۱۰۰۰۰	۵۶۹۶۰۰۰
		درصد	٪۶/۳	٪۲۶/۹	٪۶۶/۹	٪۱۰۰
	مسؤول	فراوانی	۱۴۱۰۰۰۰	۱۵۸۰۰۰۰	۱۶۸۰۰۰۰	۴۶۷۰۰۰۰
		درصد	٪۳۰/۲	٪۳۳/۸	٪۳۶/۰	٪۱۰۰
جمع کل		فراوانی	۱۷۶۶۰۰۰	۳۱۱۰۰۰۰	۵۴۹۰۰۰۰	۱۰۳۶۶۰۰۰
		درصد	٪۱۷/۰	٪۳۰/۰	٪۵۳/۰	٪۱۰۰

جدول ۲-۱۰: تعداد بازیافت‌ها برای مقوله به کاربردن "همزه" به‌صورت های مختلف (n=2)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	یاهو	آلتاویستا	
شکل واژه	مسأله	فراوانی	۱۷۴۰۰۰۰	۱۳۷۰۰۰۰	۲۰۷۰۰۰۰	۵۱۸۰۰۰۰
		درصد	٪۳۳/۶	٪۲۶/۴	٪۴۰/۰	٪۱۰۰
	مسئله	فراوانی	۱۳۰۰۰۰	۷۷۲۰۰۰	۳۶۳۰۰۰۰	۴۵۳۲۰۰۰
		درصد	٪۲/۹	٪۱۷/۰	٪۸۰/۱	٪۱۰۰
جمع کل		فراوانی	۱۸۷۰۰۰۰	۲۱۴۲۰۰۰	۵۷۰۰۰۰۰	۹۷۱۲۰۰۰
		درصد	٪۱۹/۳	٪۲۲/۱	٪۵۸/۷	٪۱۰۰

این دو کلیدواژه، از واژه‌های پربسامد در فارسی هستند و نتیجه به خوبی عدم یکدستی در نگارش "همزه" را نشان می‌دهند.

۱۱. تنوع استفاده از "الف" مقصوره

جدول۱۱: تعداد بازیافت‌ها برای مقوله تنوع استفاده از "ی" در واژه‌های عربی مختوم به "ا" (n=3)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	یاهو	آلتاویستا	
شکل واژه	موسى	فراوانی	۵۵۳۰۰۰	۵۰۶۰۰۰	۳۴۷۰۰۰۰	۴۵۲۹۰۰۰
		درصد	٪۱۲/۲	٪۱۱/۲	٪۷۶/۶	٪۱۰۰
	موسی	فراوانی	۲۴۹۰۰۰	۵۰۶۰۰۰	۳۴۸۰۰۰۰	۴۲۳۵۰۰۰
		درصد	٪۵/۹	٪۱۱/۹	٪۸۲/۲	٪۱۰۰
	موسا	فراوانی	۷۳۶	۳۳۵۰	۱۳۹۰۰	۱۷۹۸۶
		درصد	٪۴/۱	٪۱۸/۶	٪۷۷/۳	٪۱۰۰
جمع کل		فراوانی	۱۰۱۵۳۵۰	۶۹۶۳۹۰۰	۱۱۱۱۷۷۰	۹۰۹۱۰۲۰
		درصد	٪۱۱/۲	٪۷۶/۶	٪۱۲/۲	٪۱۰۰

حرف "ی"، روی بعضی صفحه‌کلیدها به شکل عربی آن نوشته می‌شود (یعنی به صورت ى). داده‌های جدول ۱۱ بیانگر این است که موتورکاوش "گوگل" برای دو شکل موسى و موسی، به طور

کامل دو نتیجه متفاوت به دست می‌دهد. این نکته باید به کاوشگران آموزش داده شود که برای نوشتن حرف "ی" به شکل فارسی، باید از کلید شیفت به همراه کلید دیگری -که در هر رایانه‌ای متفاوت است- استفاده کنند.

۱۲. تنوع املایی بعضی واژه‌ها که همه درست هستند

جدول۱۲. تعداد بازیافت‌ها برای مقوله تنوع املایی بعضی واژه‌ها که همه درست هستند (n=2)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	یاهو	آلتاویستا	
شکل واژه	اتاق	فراوانی	۱۷۴۰۰۰۰	۱۰۳۰۰۰۰	۱۰۶۰۰۰۰	۳۸۳۰۰۰۰
		درصد	٪۲۷/۷	٪۲۶/۹	٪۲۷/۷	٪۱۰۰
	اطاق	فراوانی	۱۲۵۰۰۰	۸۹۲۰۰	۹۸۷۰۰	۳۱۲۹۰۰
		درصد	٪۳۹/۹	٪۲۸/۵	٪۳۱/۵	٪۱۰۰
جمع کل		فراوانی	۱۸۶۵۰۰۰	۱۱۱۹۲۰۰	۱۱۵۸۷۰۰	۴۱۴۲۹۰۰
		درصد	٪۴۵/۰	٪۲۷/۰	٪۲۸/۰	٪۱۰۰

کاوشگری که می‌خواهد به سرعت به اطلاعات دسترسی پیدا کند، کمتر حضور ذهن دارد که تمام صورت‌های املایی را به خاطر بیاورد.

۱۳. استفاده از "ا" و "آ" به‌جای هم

جدول۱۳: تعداد بازیافت‌ها برای مقوله استفاده از "ا" و "آ" به جای هم (n=2)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	یاهو	آلتاویستا	
شکل واژه	فرآیند	فراوانی	۵۵۷۰۰۰	۴۰۳۰۰۰	۴۱۵۰۰۰	۱۳۷۵۰۰۰
		درصد	٪۴۰/۵	٪۲۹/۳	٪۳۰/۲	٪۱۰۰
	فرایند	فراوانی	۴۷۲۰۰۰	۳۶۰۰۰۰	۳۷۶۰۰۰	۱۲۰۸۰۰۰
		درصد	٪۳۹/۱	٪۲۹/۸	٪۳۱/۱	٪۱۰۰
جمع کل		فراوانی	۱۰۲۹۰۰۰	۷۶۳۰۰۰	۷۹۱۰۰۰	۲۵۸۳۰۰۰
		درصد	٪۳۹/۸	٪۲۹/۵	٪۳۰/۶	٪۱۰۰

کاوشگران از میان دو شکل بالا، آن صورتی را انتخاب می‌کنند که پیشتر به نوشتن آن عادت کرده‌اند. در نتیجه، صفحه‌های زیادی را که ممکن است حاوی اطلاعات با ارزشی باشند، از دست می-دهند.

۱۴. واژه‌های خاص در پیوسته‌نویسی و جدانویسی

جدول ۱۴: تعداد بازیافت‌ها برای دو کلیدواژه "مورچه گان" یا "مورچگان" (n=2)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	یاهو	آلتاویستا	
شکل واژه	مورچه گان	فراوانی	۶۶۲	۶۷۸	۷۰۰	۲۰۴۰
		درصد	٪۳۲/۵	٪۳۳/۲	٪۳۴/۳	٪۱۰۰
	مورچگان	فراوانی	۱۰۹۰۰	۱۶۳۰	۱۷۴۰	۱۴۲۷۰
		درصد	٪۷۶/۴	٪۱۱/۴	٪۱۲/۲	٪۱۰۰
جمع کل		فراوانی	۱۱۵۶۲	۲۳۰۸	۲۴۴۰	۱۶۳۱۰
		درصد	٪۷۰/۹	٪۱۴/۲	٪۱۵/۰	٪۱۰۰

داده‌های جدول بیانگر این است که شکل نوشتاری مورچگان بازیافت‌های بیشتری در بر داشته است و تمایل برای حذف "های بیان حرکت" در بین نویسندگان وجود دارد.

۱۵. "تای" منقوط

جدول ۱۵: تعداد بازیافت‌ها برای مقوله "تای" منقوط (n=2)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	یاهو	آلتاویستا	
شکل واژه	توراة	فراوانی	۹۳	۳۱۷	۲۱۸۰۰	۲۲۲۱۰
		درصد	٪۰/۴	٪۱/۴	٪۹۸/۲	٪۱۰۰
	تورات	فراوانی	۷۱۸۰۰	۶۵۱۰۰	۷۰۳۰۰	۲۰۷۲۰۰
		درصد	٪۳۴/۷	٪۳۱/۴	٪۳۳/۹	٪۱۰۰
جمع کل		فراوانی	۷۱۸۹۳	۶۵۴۱۷	۹۲۱۰۰	۲۲۹۴۱۰
		درصد	٪۳۱/۳	٪۲۸/۵	٪۴۰/۱	٪۱۰۰

در این مورد، تعداد کمی نتیجه برای کلمه "توراة" آمده است (به استثنای "آلتاویستا"). پس می‌توان گفت که تمایل نویسندگان به نوشتن "تای" منقوط، به صورت «ت» است. اما همچنان صفحه‌هایی وجود دارند که با "ة" ذخیره شده‌اند و ممکن است دارای اطلاعات خوبی باشند.

۱۶. نشانه "تشدید"

جدول ۱۶: تعداد بازیافت‌ها برای مقوله نشانه "تشدید" (n=2)

متغیر		آماره	موتور کاوش			جمع کل
			گوگل	یاهو	آلتاویستا	
شکل واژه	موحدی	فراوانی	۴۴۰۰۰	۴۶۷۰۰	۴۷۰۰۰	۱۳۷۷۰۰
		درصد	٪۳۲/۰	٪۳۳/۹	٪۳۴/۱	٪۱۰۰
	موحّدی	فراوانی	۲۳۰۰۰	۴۵۴۰۰	۴۸۰۰۰	۱۱۶۴۰۰
		درصد	٪۱۹/۸	٪۳۹/۰	٪۴۱/۲	٪۱۰۰
جمع کل		فراوانی	۶۷۰۰۰	۹۲۱۰۰	۹۵۰۰۰	۲۵۴۱۰۰
		درصد	٪۲۶/۴	٪۳۶/۲	٪۳۷/۴	٪۱۰۰

بیشتر برای سرعت در کار ماشین‌نویسی، "تشدید" نوشته نمی‌شود. این امر علاوه بر این که در هم‌نگاشت‌ها تولید اشکال می‌کند، طبق جدول ۱۶ موجب بازیابی نشدن صفحه‌های اینترنتی به نسبت یکسان با کلیدواژه تشدیددار می‌شود.

سؤال ۳. آیا رابطه معناداری بین شکل واژه‌ها و نوع ابزار جستجو وجود دارد؟

میزان χ^2 مشاهده شده در سطح معناداری "α=۰/۰۵" نشان می‌دهد که بین شکل واژه و ابزار جستجو رابطه معناداری وجود دارد. بنابراین می‌توان نتیجه گرفت که به کار بردن یک شکل خاص از کلیدواژه و نیز استفاده از یک ابزار جستجوی خاص، در بازیابی اطلاعات اثرگذار است. برای نمونه جستجوی واژه"مسئله" در موتور کاوش "آلتاویستا" بازیافت بیشتری نسبت به جستجوی واژه "مسأله" دارد، اما جستجو در موتور جستجوی یاهو به طور کامل نتیجه‌ای متفاوت به دست می‌دهد.

همچنین مقدار ضریب Phi به دست‌آمده بیانگر میزان ارتباط بین دو متغیر است. ضریب "فی"، معیاری برای ارزیابی هم‌آیندی بین دو متغیر است. دامنه مقدارهای این معیار که جهت هم‌آیندی را هم نشان می‌دهد، از ۱- تا ۱+ است. از این معیار می‌توان در ارزیابی هم‌آیندی میان متغیرهای اسمی چندین

مقوله‌ای استفاده کرد (کورتز، ۱۳۷۴ :۳۴۵).

جدول ۱۷: آزمون خی دو برای بررسی رابطه معناداری بین شکل واژه و ابزار جستجو

متغیر	درجه آزادی	χ^2	سطح معناداری	ضریب فی
پتاسیم/پتاسیوم	۲	۱۰۹۸۴۶۵	۰/۰۰۱	۰/۰۸۴
زبانشناس/ زبان شناس/ زبان‌شناس	۳	۲۰۴۳۵۸۶۴	۰/۰۰۱	۰/۵۰۲
دقیقاً/ دقیقن	۲	۲۳۸۰۶۷۱۰	۰/۰۰۱	۰/۱۰۰
شمشیر باز/ شمشیرِ باز	۲	۲۵۹۹۵۴۸۳	۰/۰۰۱	۰/۱۷۵
پرتو آفتاب/ پرتوی آفتاب	۲	۱۲۱۵۶۳۹۳	۰/۰۰۱	۰/۲۲۴
باغ ها/ باغها	۲	۱۱۲۴۰۹۲۰	۰/۰۰۱	۰/۰۶۵
خانه من/ خانهٔ من/ خانه‌ی من	۳	۷۹۰۲۹۲۵	۰/۰۰۱	۰/۵۰۲
املا/ املاء	۲	۱۴۰۴۴۰۸۴	۰/۰۰۱	۰/۱۳۷
عطایی/ عطائی	۲	۳۳۰۸۷۲۶۰	۰/۰۰۱	۰/۲۵۴
مسؤول/ مسئول	۲	۱۹۴۵۴۷۰۹۵۰	۰/۰۰۱	۰/۴۴۸
مسئله/ مسأله	۲	۱۹۴۵۴۷۰۹۵۰	۰/۰۰۱	۰/۳۶۳
موسی/ موسی/ موسا	۳	۱۰۶۸۹۹۷۰۴	۰/۰۰۱	۰/۱۱۰
اتاق/ اطاق	۲	۳۷۵۷۱۰۴	۰/۰۰۱	۰/۰۳۰
فرایند/ فرآیند	۲	۵۷۲۸۵۱	۰/۰۰۱	۰/۰۱۵
مورچه گان/ مورچگان	۲	۱۶۶۹۸۶۱	۰/۰۰۱	۰/۳۲۰
تورات/ توراة	۲	۳۴۴۴۰۱۴۸	۰/۰۰۱	۰/۳۸۷
موحدی/ موحّدی	۲	۴۸۵۹۶۳۴	۰/۰۰۱	۰/۱۳۸

بحث و نتیجه‌گیری

۱. شیوه نگارش فارسی باعث بروز چالش های جدی در امر نمایه سازی این زبان شده است. در بسیاری موارد چند شکل نگارشی، برای یک واژه، درست شمرده شده است. این چندگونگی شکل واژه‌ها، برای رایانه قابل درک نیست. چراکه رایانه واژه‌ها را تنها به همان صورتی که ذخیره کرده است می‌شناسد و بازیابی می‌کند. بنابراین در مقابل سایر شکل‌های نوشتاری، آن را اصطلاح دیگری محسوب کرده و در هنگام جستجوی اطلاعات آن را بازیابی نمی‌کند.

۲. هیچکدام از موتورهای کاوش، چالش‌های شیوه‌های نگارش فارسی را به منظور بهبود نتیجه کاوش، مورد توجه قرار نداده‌اند. همان‌طورکه سروکا (Sroka, 2000)، بارایلان و گتمان (Bar-Ilan & Gutman, 2002)، مقداد (Moukdad, 2005) و تاث (Toth, 2006) در بررسی های خود نشان دادند، موتورهای کاوش عمومی وب، ویژگی‌های زبانی زبان‌های غیر انگلیسی را در جستجوی اطلاعات نادیده می‌گیرند. این یافته‌ها نگران کننده است، زیرا این موتورهای کاوش (برای مثال "گوگل") در کشور ما، بسیار مورد توجه می‌باشند و بیشتر کاربران، از آنچه به هنگام جستجو در این ابزارها از دست می‌دهند، آگاهی ندارند.

۳. *بین شکل واژه و ابزار جستجو رابطه معناداری وجود دارد.*

بنابراین، به کار بردن یک شکل خاص از کلیدواژه و نیز استفاده از یک ابزار جستجوی خاص، در بازیابی اطلاعات اثرگذار است.

ابزارهای کاوش اینترنت، مهمترین فناوری حاضر برای دسترسی به اطلاعات در محیط وب به شمار می‌آیند. در حال حاضر ابزارهای کاوش مختلفی در جهان ظهور پیدا کرده‌اند. لیکن ابزارهای جستجویی که امکان جستجوی اطلاعات به زبان فارسی را ارائه می‌دهند، محدود می‌باشند. از طرف دیگر، امکانات و قابلیت‌های آنها برای بازیابی کارآمد و مناسب اطلاعات متفاوت است. بی شک ایجاد یک ابزار کاوش قوی ملی، تحت نظارت سازمان‌های رایانه‌ای و انجمن‌های زبان‌شناسی و منطبق با نیازهای اطلاعاتی کاربران اینترنت در ایران می‌تواند میزان دقت در کاوش اطلاعات را بالا ببرد.

از طرفی به دلیل وجود مشکل هایی در خط فارسی، همواره برای استفاده از نرم‌افزارهایی چون OCR[1] برای وارد کردن متن توسط پویشگر[2] به رایانه و همچنین استفاده از واژه پردازها و ابزارهای مورد استفاده در آنها مانند غلط یاب‌های دستوری و املایی محدودیت هایی وجود دارد. پردازش، خلاصه سازی و بازیابی اطلاعات ازمحتوای متن، تحلیل آن و استفاده از نرم‌افزارهای تبدیل متن به گفتار و برعکس، همه از مواردی هستند که به دلیل محدودیت‌های خاص خط فارسی، استفاده از آنها به صورت کامل انجام نمی پذیرد؛ پایگاه‌های اطلاعاتی که با استفاده از شیوه خط کنونی به ذخیره و بازیابی اطلاعات می‌پردازند، نمی‌توانند کارایی مطلوب داشته باشند. همچنین عواملی که بدان اشاره شد، سبب کندی مراحل ذخیره و بازیابی اطلاعات شده و به نسبت بازیافت اطلاعات را کاهش می‌دهند. پایگاه‌های اطلاعاتی مدارک

[1] Optimal Character Recognizer
[2] Scaner

فارسی با وجود عمر کوتاه‌شان با مشکل های بسیاری درگیر هستند. در صورتی که برای این مشکل‌ها چاره‌ای اندیشیده نشود، با توجه به حجم فزاینده اطلاعات، مهار آنها آسان نخواهد بود.

پیشنهادهایی برای رفع چالش

۱. بیشتر واژه‌پردازهای لاتین، به نظام غلط‌یاب املایی مجهز هستند که کار تصحیح متن الکترونیکی را به صورت خودکار انجام می‌دهد. پیشنهاد می‌شود روش و الگوریتم‌های غلط‌یاب فارسی در برنامه پژوهشی استادان زبان‌شناس فارسی قرار گیرد تا تولیدکنندگان نرم افزارهای واژه‌پرداز فارسی بتوانند با به کارگیری این روش‌ها، نرم‌افزارهای خود را به غلط‌یاب املایی خودکار مجهز کنند.

۲. در واژه‌پردازهای پیشرفته با توجه به اصول دستور زبان، امکان تصحیح دستوری متن الکترونیکی گنجانیده شود. برای دست‌یابی به نظام مشابه در زبان فارسی، لازم است استانداردهای نوشتار فارسی برای رایانه تدوین شود.

۳. نظام‌های بازیابی اطلاعات در زبان انگلیسی از امکانات ریشه‌سازی استفاده زیادی می‌کنند. در اینگونه نظام‌ها، با وارد کردن یک واژه به عنوان کلیدواژه، نظام به‌طور خودکار تمامی مشتقات واژه را نیز جستجو می‌کند. برای نمونه اگر جستجوی "کتاب" مدنظر ما باشد، واژه‌هایی نظیر"کتابخانه"، "کتابداری"، "کتاب فروش" و مانند آنها نیز بازیابی می‌شود. پورتر[1] یکی از توانمندترین نظام‌های ریشه-یابی در زبان انگلیسی است. این نظام بر پایه دسته‌بندی واژه‌ها به کمک واج‌ها و هجاها بنا نهاده شده‌است. برای ایجاد چنین نظامی در زبان فارسی، باید متخصصان زبان‌شناسی و رایانه همکاری نزدیکی با هم داشته‌باشند.

۴. در جهان کنونی بازنگری در شیوه نگارش فارسی را باید به شکل متفاوتی نسبت به گذشته انجام داد. اگر در گذشته ادیبان به تنهایی برای این امر تصمیم می‌گرفتند، هم اکنون تمامی کسانی که به نحوی با خط سر و کار دارند باید در تصمیم‌گیری دخالت داشته باشند. "فرهنگستان زبان و ادب فارسی"، می‌تواند افرادی شامل گروه‌های زیر را مامور این کار کند:

- نویسندگان، شاعران، مترجمان، روزنامه‌نگاران و تمامی افرادی که به کار نوشتن مشغول هستند.

[1] Porter

- ویراستاران ونسخه‌پردازان؛ که تا کنون بار یکدست کردن شیوه نگارش متون بر دوش آنها بوده است.
- زبان‌شناسان؛ که پیشنهادها را با توجه به ساختار و کاربرد زبان، ارزیابی کنند.
- خوشنویسان و طراحان و به ویژه طراحان حروف؛ در طرح‌های موجود احساس می‌شود که توجه خاصی به زیبایی‌شناسی خط فارسی نشده است، در حالی که توجه به این امر، که بخشی از میراث فرهنگی و هویت ملی ما به شمار می‌آید کاری الزامی به نظر می‌رسد.
- برنامه‌نویسان رایانه؛ که راهگشای قابلیت های بی‌شمار رایانه برای فارسی‌زبان‌ها خواهند بود.
- کتابداران و اطلاع‌رسانان، که کار ترجمه نیازهای اطلاعاتی کاربران به زبان رایانه بر عهده آنها است و به طور مستقیم با این چالش‌ها سر و کار دارند.

کتابنامه

اسلامی، محرم (۱۳۸۱). دشواری‌های پردازش رایانه‌ای خط فارسی. *نشر دانش*، ۱۹(۳).

فرهنگستان زبان و ادب فارسی(۱۳۸۳). *دستور خط فارسی*. تهران: فرهنگستان زبان و ادب فارسی.

حری، عباس (۱۳۷۲). کامپیوتر و شیوه نگارش فارسی، *پیام کتابخانه*، ۳(۱).

راثی ساربانقلی، محمد صابر (۱۳۸۴). *بررسی مشکلات جستجو و بازیابی اطلاعات به زبان فارسی از اینترنت با مطالعه موردی بر روی کاربران مرکز اینتـرنت دانشگاه آزاد اسلامی واحد شبستر*. پایان نامه کارشناسی ارشد کتابداری و اطلاع‌رسانی، دانشگاه آزاد اسلامی، واحد تهران شمال.

سرمد، زهره (۱۳۷۶). *روش‌های تحقیق در علوم رفتاری*. تهران: آگه.

صدیق بهزادی، ماندانا (۱۳۷۱). ناهماهنگی ضبط نام‌های بیگانه در فارسی. *فرهنگ*،۱۳.

صدیقی، محسن (۱۳۸۳). روشی برای رفع چالشهای محتوا کاوی وب های فارسی. *نما*، ۴(۲). بازیابی در ۱۳۸۷/۱۲/۱۰ از http://www.irandoc.ac.ir/data/e_j/vol4/shahidi.htm.

کابلی، ایرج (۱۳۷۳). پیوسته‌نویسی، جدانویسی یا بی‌فاصله‌نویسی. *آدینه*، (۹۶).

کورتز، نورمن (۱۳۷۴). *مقدمه‌ای بر آمار در علوم اجتماعی*. (ترجمه حبیب الله تیموری). تهران: نشر نی.

کوشا، کیوان (۱۳۸۱). معیارهای ارزیابی ابزارهای کاوش اینترنت: مطالعه مقایسه‌ای‌بر روی ابزارهای کاوش وب با واسط جستجوی فارسی. *نشریه الکترونیکی کتابدار*. بازیابی در ۱۳۸۷/۱۲/۱۰ از http://www.ketabdar.org/magazine/detailarticle.asp?number=25.

مرتضایی، لیلا (۱۳۷۶). مسائل زبان و خط فارسی در ذخیره‌سازی و بازیابی اطلاعات. *فصلنامه اطلاع رسانی*. ۱۷(۱-۲).

منصورفر، کریم (۱۳۷۴). *روش‌های آماری*. تهران: دانشگاه تهران. موسسه چاپ و انتشارات.

یوسفان نجف آبادی، احمد (۱۳۸۲). یک سیتم *بازیابی اطلاعات متنی برای زبان فارسی بر پایه نمایه‌گذاری معانی پنهان*. پایان نامه کارشناسی ارشد مهندسی کامپیوتر، دانشگاه شیراز.

Bar-Ilan, J. & Gutman, T. (2002). How Do Search Engines Handle Non-English Queries?: A Case Study Retrieved February 2, 2009 from Sciencedirect database.

Hedlund, T et al. (2000). Aspects of Swedish Morphology and Semantics from the Perspective of Mono- and Cross-Language Information Retrieval. Retrieved January 23, 2009 from Elsevier database.

Monz, C. & De Rijke, M. (2002). Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German, and Italian. In proceedings of *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation forum, CLEF 2001*. September 3-4 2001. Darmstadt, Germany.

Moukdad, H. (2005). Lost in Cyberspace: How Do Search Engines Handle Arabic Queries?. *The International Information & Library Review*, 37 (4), 237-394.

Sroka, M. (2000). Web Search Engines for Polish Information Retrieval: Questions of Search Capabilities and Retrieval Performance. *Library Review*, (32), 87-98.

Toth, E. (2006). Exploring the Capabilities of English and Hungarian Search English for Various Queries. *Library Review*, (56), 38-47.